A Network-Based Model for Predicting Air Traffic Delays

Juan Jose Rebollo and Hamsa Balakrishnan
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, MA, USA
Email: {jrebollo,hamsa}@mit.edu

***Abstract*—This paper presents a new model for predicting delays in the National Airspace System (NAS). The proposed model uses Random Forest (RF) algorithms, considering both temporal and spatial (that is, network) delay states as explanatory variables. In addition to local delay variables that describe the arrival or departure delay states of the most influential airports and origin-destination (OD) pairs in the network, we propose new network delay variables that depict the global delay state of the entire NAS at the time of prediction. The paper analyzes both the classification and regression performance of the proposed prediction models, which are trained and validated on 2007 and 2008 ASPM data. The predictive performance of the model is evaluated using the 100 most delayed OD pairs in the NAS: the results show that given a 2-hour prediction horizon, the average test error across these 100 OD pairs is 19% when classifying delays as above or below 60 min. The effect of changes in the classification threshold and prediction horizon on model performance are also studied.**

***Index Terms*—Air traffic delay prediction; network effects; k-means clustering; Random Forest methods; ASPM**

I. Introduction

The large number of shared resources in the air traffic network together with aircraft, crew and passenger interdependencies makes air traffic network effects an important field of study [1, 2]. Network effects are becoming more significant for two main reasons. Firstly, airlines attempt to maximize aircraft utilization in order to increase revenue, thereby reducing the time buffer between arrivals and departures. As a result, arrival delays are more likely to be propagated to subsequent departure flights [1]. Secondly, as demand gets closer to the maximum capacity levels, the ability of the network to absorb disruptions decreases, thereby making the network susceptible to large-scale delays. The study of network effects can help us to understand factors that mitigate or amplify delay propagation, and to identify the elements of the network causing a bigger impact on the entire system.

The goal of this paper is to study the potential of delay interdependencies in the NAS network in developing delay prediction models. Can we predict the departure delay of a particular OD pair by only looking at the current and/or past delay state of the different elements in the network? Similarly, we hypothesize that the delay state of the different elements in the network at a certain time would be a good indicator of how NAS delays will evolve in the short term. We expect that our prediction models will have difficulties capturing non-congestion related delays, which only affect a few elements in the network (for example, delays related to mechanical issues which only affect a small subset of flights). Our goal is not to predict individual flights delays, but instead to estimate the future network-related delay on a certain route. We evaluate the prediction performance of this model over actual delay data, which can include any type of delay. While different prediction models have been proposed in the research community [3-8], none of these models have investigated the role of the network delay state.

We consider three different types of variables in our delay prediction models. First, we have temporal variables, which only depend on the time for which the prediction is being made (for example, the time of day, or day of week), and not the delay state of the network. Second, we have local delay state variables, which indicate the delay level of specific elements of the network (for example, delay at a particular airport or route). Finally, we have high-level delay state variables which depict the state of a group of elements, and are obtained by clustering local delay state variables.

The analysis of the different prediction models presented in this paper will help us better understand delay interactions among the different elements in the NAS network, and evaluate how much of the future delay on a particular route can be explained by looking at the current network delay state.

The paper is organized as follows. Section II describes the data used in this research, and the preprocessing performed. Section III analyzes the explanatory variables that will be used in the prediction models. Sections IV and V focus on the prediction model description, and performance analysis. Finally, the paper ends with conclusions and next steps.

II. Problem Statement

The main objective of this paper is to predict the departure delay state of a certain route in the network. The departure delay state at time t is an estimate of the departure delay of flights taking off at time t in that route. We evaluate the value of different network delay state variables in predicting the departure delay state of a specific route. We study two types of prediction mechanisms: *classification*, where the output is a binary prediction of whether the delay is more or less than a predefined threshold, and *regression*, where the continuous output is an estimate of the delay along the route of interest.

III. Input data and preprocessing

The results presented in this paper were obtained using data from the FAA's Aviation System Performance Metrics (ASPM) database. The ASPM database integrates data from different sources: ETMS, ARINC, OAG and ASQP. ASPM

provides detailed data for individual flights by phase of flight, airport weather data, runway configuration, and arrival and departure rates. Two years of ASPM data were processed in our analysis, from January 2007 to December 2008. We processed the following ASPM fields for each flight:

- Dep_LOCID: Departure Location Identifier.
- Arr_LOCID: Arrival Location Identifier.
- SchInSec: Scheduled Gate-In.
- ActInSec: Actual Gate-In.
- SchOffSec: Scheduled Wheels-off.
- ActOffSec: Actual Wheels-off.
- FAACARRIER: Flight Carrier Code
- TAILNO: Aircraft Tail Number

These ASPM fields correspond to individual flight data, which are processed to obtain a more robust aggregate delay picture. We are not interested in predicting individual flight delays, but instead in the delay levels of different airports and OD pairs in the network. We define the delay state of an airport or OD pair at time t as an estimate of the delay that a hypothetical flight using that resource at that time will experience. For example, if the BOS-MCO departure delay state is 30 min at 3 pm, it means that the estimated departure delay for a BOS-MCO flight taking off at 3pm is 30 min.

We use a moving median filter to obtain the delay states of airports and OD pairs. The delay state of any NAS element at time t refers to the median delay of all the flights that fall within a window of size W centered at time t. This low pass filter mitigates high frequency changes by calculating the median of the data points. The window size is set to two hours, and the time step to one hour.

Finally, the two years of data led to 2,029 airports, and 31,905 origin destination pairs. Most of these links average less than one flight a day. Since only links with high traffic volume can have an impact on the rest of the network, only OD pairs with 10 or more flights per day are included in the analysis. Figure 1 depicts the resulting simplified network, which is composed of 584 OD pairs, and 112 airports.

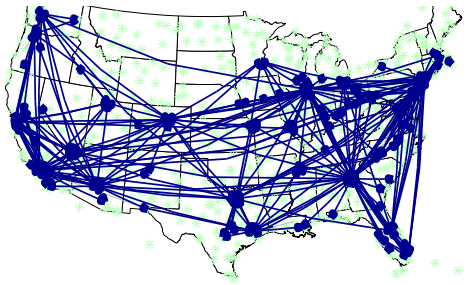

Fig. 1: Simplified NAS network showing links with at least 10 flights a day. The light green icons denote airports in the original dataset that are not included in the simplified network.

IV. Analysis of Explanatory Variables

This section describes the explanatory variables that are included in the delay prediction model, and demonstrates their relevance. The analysis of the different variables presented here focuses on the JFK-ORD departure delay prediction model. However, since our goal is to predict delays along arbitrary links in the network, the explanatory variables are defined for generic OD pairs.

The Kruskal-Wallis parametric ANOVA test [9] and the multiple comparisons test were used to evaluate the dependence of the future departure delay with different categories for the proposed categorical explanatory variables. A parametric ANOVA test was used due to the highly skewed delay distributions. By contrast, a Random Forest (RF) methodology was used to identify the most relevant continuous variables. Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. The method combines bagging idea and the random selection of variables at each tree split.

The significance of the explanatory variables depends on the output of the prediction model considered (e.g., time horizon, regression vs. classification, etc.). We assume that if an explanatory variable has an effect on the continuous delay output, it will also have an effect on the binary output. For this reason, only the continuous delay output is analyzed in this section. While the results shown here were obtained for a 2-hour prediction window, the effect of the prediction horizon on model performance is discussed in Section V-E.

A. Temporal variables

Temporal variables considered include the time-of-day, day-of-week, and month-of-year. All these variables are categorical (e.g., the time-of-day variable has one category for each hour). The low p-values obtained by the ANOVA tests showed that the three temporal categorical variables lead to significant differences in the output delay. For example, Figure 2 presents the multiple comparisons test plot for the JFK-ORD departure delay model, showing that the confidence intervals do not overlap for most categories. Figure 2 indicates that flights departing at 3 am are the most delayed, which is reasonable, since few flights are scheduled for that hour, and any flights that actually depart at that time were likely scheduled for much earlier in the day. We also see that delays tend to accumulate through the day until demand levels drop overnight, causing delays to decrease.

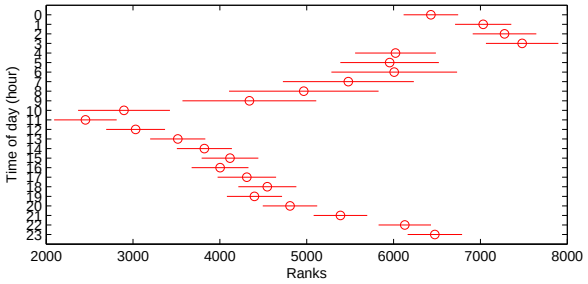

Fig. 2: Time-of-day multiple comparisons test for JFK-ORD departure delays.

B. Network delay state variables

1) NAS delay state: The NAS delay state is a categorical variable that depicts the level of delay in the entire NAS. The NAS delay state at time t is defined by the departure delay state of each link in the simplified network at time t. The typical NAS delay states are obtained by clustering the NAS delay states into N groups using the k-means algorithm. The output of the clustering algorithm will indicate the closest typical state to each of the observations, where the "typical states" are given by the centroids of each of the clusters.

The first step is to select the number of clusters, or typical NAS delay states. Six clusters are chosen for two reasons: First, the total intra-cluster distance does not decrease much for more than 6 clusters (as seen in Figure 3), and second, six appears to be qualitatively reasonable since all the main delay centers are represented in the centroids of clusters.

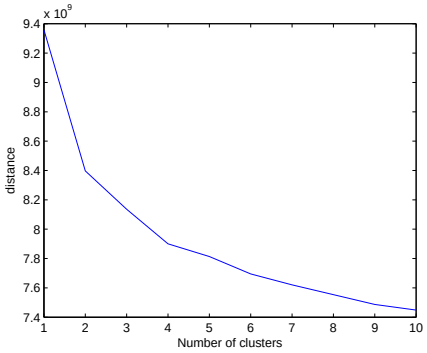

Fig. 3: Total inter-cluster distance vs. number of clusters.

Figures 4 depicts the cluster centroids' delay levels, and we can see that Chicago, New York City, and Atlanta are the main delay centers. For fewer than six clusters, Atlanta does not appear; for more than six clusters, no new delay centers appear. For simplicity, we want to have the minimum number of clusters that allows us to differentiate between all the typical delay situations.

The NAS delay state categorical variable represents the closest typical NAS delay state at that time. We are interested in evaluating the dependence of the future delay of a given OD pair on the current delay state of the NAS as a whole.

For the JFK-ORD OD pair, we performed an ANOVA test and obtained a p-value equal to zero, meaning that the means of the JFK-ORD future departure delay for different values of the NAS state categorical variable are not equal. Figure 5 shows the associated multiple comparisons intervals. It is reasonable that State 4 leads to the lowest delay interval, since it is the low NAS delay state. The next highest JFK-ORD delays are for State 1, which is the medium NAS delay state. The ATL high delay state (6) comes next: The JFK-ORD delay levels are not too high for this state. The next state is the NYC medium-high delay one (State 2), and finally we have the Chicago and NYC high delay states (3 and 5).

While some of the NAS delay state categories could be merged in this case, we do not want to make model simplifications that could worsen the model performance on other OD pairs. For example, the differences between States 1 and 6 increase significantly for the ATL-MCO route.

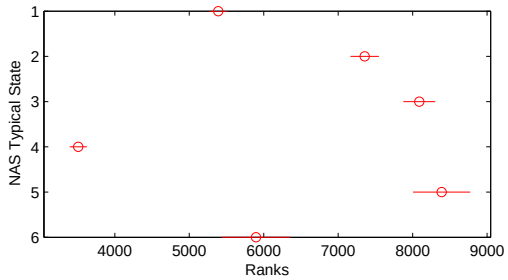

Fig. 5: NAS typical states multiple comparisons test.

2) NAS type of day: In addition to clustering the NAS delay state at time t, we clustered entire days. The idea was to identify a set of typical NAS day types, according to the daily delay of all the links in the simplified network. We hypothesized that the type of NAS day would have an impact on the future delay level of any given OD pair.

We followed the same methodology presented in the previous section to choose the number of clusters (based on distance reduction and qualitative description of the centroids), and we chose six again as the number of clusters. Since we need a video to visualize the cluster centroids, Table I describes the main source of delay at the highest delay point of the day for each of the clusters. The average daily delay is also shown.

TABLE I
TYPE OF NAS DAY CLUSTERING. DELAY DEFINITIONS: HIGH (90 MIN), MEDIUM-HIGH (60 MIN), MEDIUM (20 MIN), LOW (5 MIN).

	Avg. delay (min)	**Qualitative Description**
Day 1	29	NYC high+, ATL, ORD high delay
Day 2	22	CHI high, NYC medium high delay
Day 3	15	NYC, ORD medium delay
Day 4	21	ATL high, NYC, ORD medium high delay
Day 5	9	Low NAS delay
Day 6	19	NYC high, ATL, ORD medium delay

Figure 6 shows the monthly occurrences of each type of day. We see that Day 1 (high NYC delays, and significant ORD and ATL delays) is more common in the summer months, while Day 6 (high NYC delays, but not high ORD or ATL delays) is seen year-round, with higher frequency around the summer months. We also see that the Chicago high delay day (Day 2) is more frequent in winter, while the Atlanta high delay day (Day 4) is more frequent in summer.

Finally, in Figure 7 presents the multiple comparisons test results for the JFK-ORD departure delay and the type-of-day variable, showing different JFK-ORD departure delay levels for different categories of the type-of-day variable.

We note that one needs the entire day's delay information to determine the type of a given day. In practice, if we make a delay prediction at 2 pm, we only have the delay information from the beginning of the day to 2 pm. Although the type of day should be estimated with the information available at the time of prediction is made, we are going to assume that we know the type of day with certainty before the day is over.

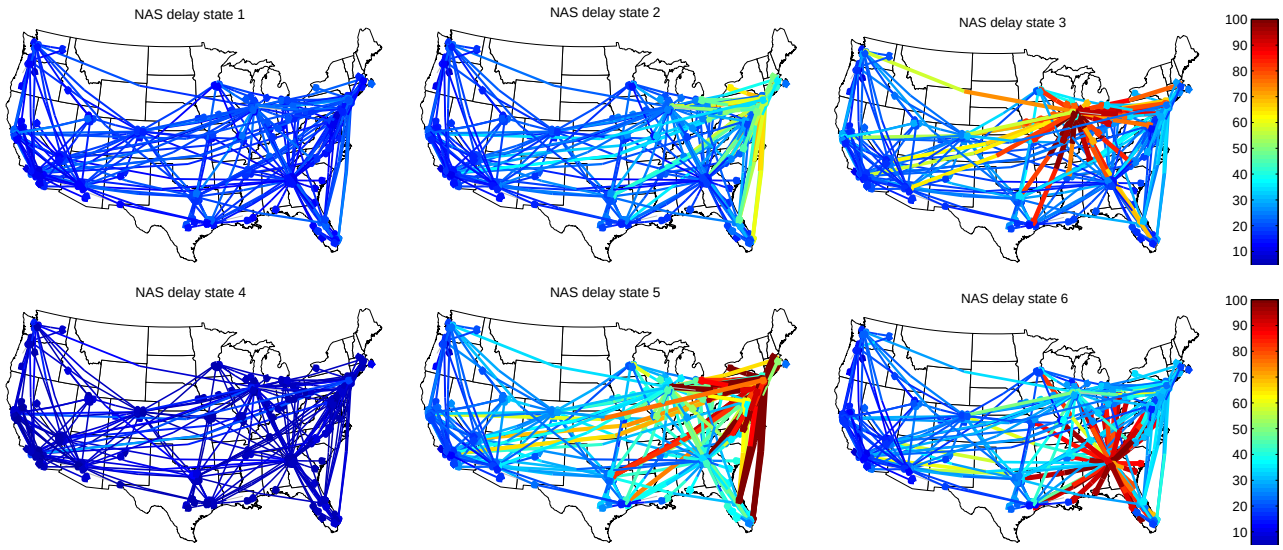

Fig. 4: Centroids of NAS delay states for six clusters.

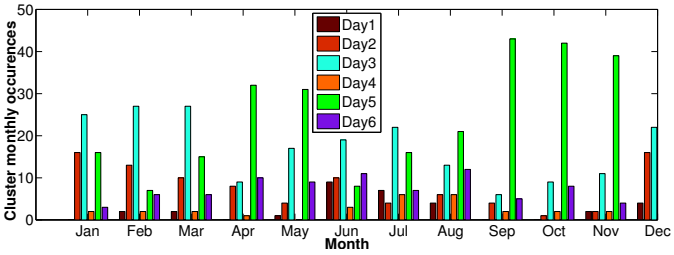

Fig. 6: Monthly occurrences of NAS type-of-day.

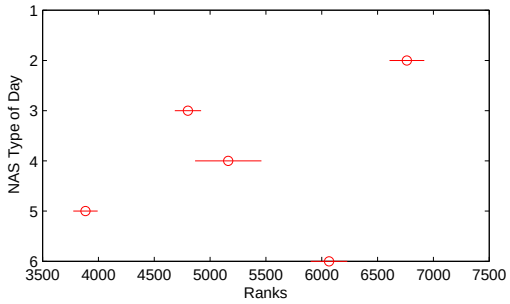

Fig. 7: Type of NAS day multiple comparisons test.

While evaluating the prediction capabilities of the type-of-day variable, we do not include the errors in estimating it.

NAS delays for the previous day are known with certainty, and can help predict delays later in the day. The NAS does not immediately recover from high delay situations, such as, a day with strong convective weather and a large number of cancelled flights. Passengers will be accommodated in flights over the next few days, leading to higher traffic levels and subsequent delays. Scheduled aircraft routings are also affected by cancelled flights, causing additional delays. The multiple comparisons test results showed significant differences for the previous day type variable, and this variable was therefore included in the prediction model.

3) Influential airport delay states: The influential airports for a given delay prediction problem are those airports whose arrival or departure delay states play an important role in predicting the delay of the OD pair of interest. For example, while predicting JFK-ORD departure delays, it is reasonable that the ORD departure delay and the JFK arrival delay will play an important role. We are interesting in identifying other airport delay variables that would improve the predictions.

We consider 400 possible airports, which led to 800 variables (departure and arrival delay treated independently). The goal was to order the 800 variables according to their prediction capabilities. For this purpose we generated a Random Forest [10] with the 800 airport delay variables as explanatory variables, and the delay to be predicted as the output. We used the variables' importance provided by the RF algorithm to choose the most relevant airport delay variables. Table II shows the selected 10 variables for the JFK-ORD departure delay prediction model.

TABLE II
INFLUENTIAL AIRPORTS FOR JFK-ORD DEPARTURE DELAY PREDICTION.

Airport	Delay Type	Variable Importance
DCA	Departure	100
JFK	Departure	96.9
ORD	Arrival	85.3
ORD	Departure	82.8
LGA	Departure	58.9
BOS	Departure	58.9
PHL	Departure	58.2
EWR	Departure	57.7
JFK	Departure	56.3
DCA	Arrival	46.1

4) Influential OD pair delay states: Our goal was to identify the OD pairs whose arrival/departure delays can have an important role in a delay prediction model. We used the same methodology presented in the previous section, but with

OD pairs delay variables instead of airport delay variables. We included in our analysis all the OD pairs in the simplified network. This led to 1,064 variables, half of which were arrival delay variables, and the other half, departure delay variables.

For the JFK-ORD model, the RF algorithm identified the 10 most important OD pairs presented in Table III. Intuitively,

TABLE III
INFLUENTIAL OD PAIRS FOR JFK-ORD DEPARTURE DELAY PREDICTION.

Origin	Destination	Delay Type	Variable Importance
JFK	ORD	Departure	100
EWR	ORD	Departure	90.9
LGA	ORD	Departure	65.3
ORD	JFK	Departure	44
ORD	LGA	Departure	24.3
BOS	ORD	Departure	17
PHL	ORD	Departure	16.9
JFK	FLL	Departure	11.9
BUF	JFK	Arrival	11.4
LGA	ORD	Arrival	11

the findings of Table III are reasonable, since most of the important variables reflect the delays prevalent in the NYC and ORD areas. There are, however, some interesting findings, for example, that the JFK-FLL departure delay has the same importance as the BUF-JFK and LGA-ORD arrival delays when predicting the JFK-ORD departure delay.

V. DELAY PREDICTION MODELS

First, we describe the training and test sets that were used to fit and test the predictive models. We sampled 10 training sets (3,000 points each) and 10 test sets (1,000 points each) from the 2007-2008 data set. We fit and tested the prediction models for each of the 10 training and test set pairs. This allowed us to obtain a measure of the error variability and a good estimate of the test error. The training and test sets were not randomly sampled from the 2007-2008 data; instead we used over-sampling. Over-sampling is the "deliberate selection of individuals of a rare type in order to obtain reasonably precise estimates of the properties of this type. In a population which includes such a rare type, a random sample of the entire population might result in very few (or none) of these individuals being selected" [11]. Over-sampling allows us to have a balanced data set, and to therefore avoid having more low delay data points in our training and test sets. This is especially important in the classification problem: If we want to classify future delays as high (e.g., over 60 min) or low (under 60 min), we want half of the points in our training and test sets to present delays of over (or under) 60 min.

We tested different classification and regression models (logistic regression, single classification trees, bagging, boosting, linear regression, neural nets), and the RF algorithm showed the best performance. All the results presented in the rest of the paper were obtained for the RF prediction model.

VI. DEPARTURE DELAY PREDICTION FOR THE 100 MOST-DELAYED OD PAIRS

With the purpose of evaluating our prediction model performance, we test the RF prediction model on 100 different OD pairs. We selected the 100 OD pairs with the highest average delay to avoid a shortage of high delay data points. In this section, we study the performance of the classification-based and regression-based departure delay prediction models for a 2-hour prediction window and a 60 min classification threshold (that is, whether the delay will be above or below 60 min).

A. Classification performance

We first study the classification performance. Figure 8 shows the test error histogram for the 100 most delayed OD pairs. The test error ranges from 11.3% to 28.8%, and the average value is 19.1%. The link with the lowest test error is EWR-ATL (11.3%), and the one with the highest is LAS-SFO (28.8%). Delays for flights arriving or departing from SFO are hard to predict: The average test error rate for links that have SFO as origin or destination is 23.3%. We find that 90% of the analyzed links have a test error standard deviation under 1.7 percentage points. The empirical cdf of the test error standard deviations of all the links in presented in Figure 9.

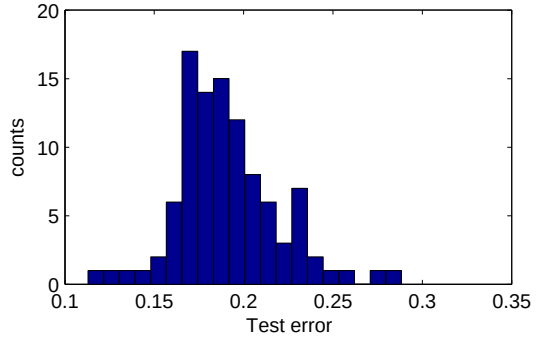

Fig. 8: Classification test error histogram for the 100 most-delayed OD pairs.

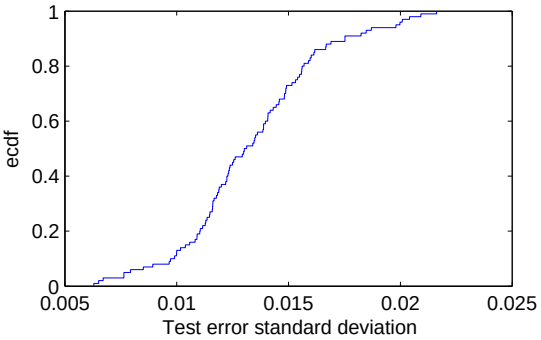

Fig. 9: Empirical cdf of the standard deviation of the classification test error for the 100 most-delayed OD pairs.

If we break down the test error into false positive and false negative error rates (FPR and FNR respectively) we see that the FNR is clearly dominant. For the 100 most delayed links, the average FNR is 23.62% and the average FPR is 14.6%, and the FNR rate is higher than the FPR for all OD pairs. In other words, the classifier is more likely to miss a high delay link than to predict high delay when in reality the delay on the

OD pair is low. This is because our prediction model bases its predictions on the delay state of the different elements in the network, and therefore has trouble capturing local delay causes (such as, mechanical issues). If delays in the relevant network elements are high, we will likely have a high delay situation in 2 hours in our link of interest; however, if the network delay is low we can still have a high delay in 2 hours due to a local issue that only affects a certain flight. The analysis also shows an increase of the FNR dominance as the test error increases. Figure 10 shows the FPR and FNR versus the test error for all the studied OD pairs. We see that the separation among the FP points and FN points increases as the test error increases. For the lowest test error OD pair, FNR/FPR ratio is 1.3, while for the highest test error FNR/FPR=1.9, showing that the FNR dominance increases with the test error. In the OD pair with the highest test error (LAS-SFO), the prediction model misclassifies high delay points almost twice as often as the low delay points.

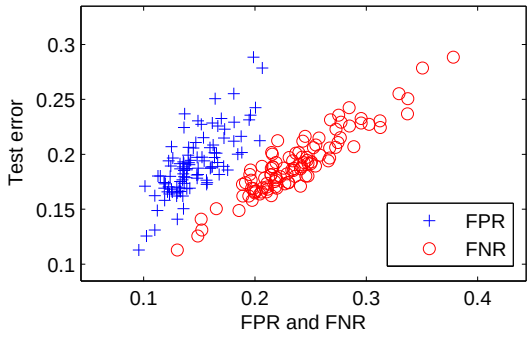

Fig. 10: FNR and FPR scatter plot for the 100 most-delayed OD pairs.

The time-of-day explanatory variable is the most important variable for both the OD pair presenting the lowest test error (EWR-ATL), and the highest (LAS-SFO). The differences in their performances can be explained using Figures 11 and 12. They show the EWR-ATL and LAS-SFO departure delay means and one standard deviation confidence intervals versus the time-of-day for the data points in the test set. We see that the EWR-ATL confidence intervals overlap less with the 60 minute threshold line than the LAS-SFO intervals. The more the overlap and the less the distance from the intervals' center to the 60 min threshold, the worse the prediction performance, because the difference between the likelihood of being over and under the decision threshold at a certain time decreases (we move towards the random guess). The LAS-SFO confidence intervals in Figure 12 are wider than the EWR-ATL intervals. This indicates less correlation between the departure delay and the time-of-day variable, and it increases the overlap with the threshold line.

B. Regression performance

Next, we take a look at the regression problem, and compare its performance with the results obtained for classification. We use the same data set as the one used in Section VI-A.

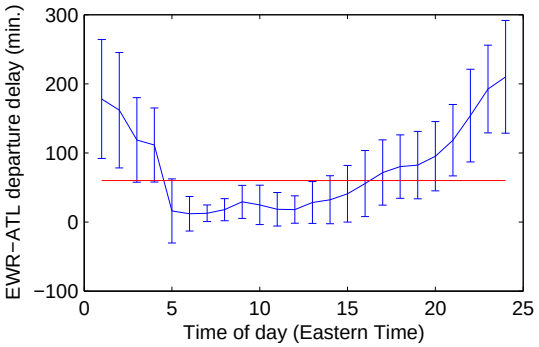

Fig. 11: EWR-ATL mean delay by time-of-day ($\pm\sigma$).

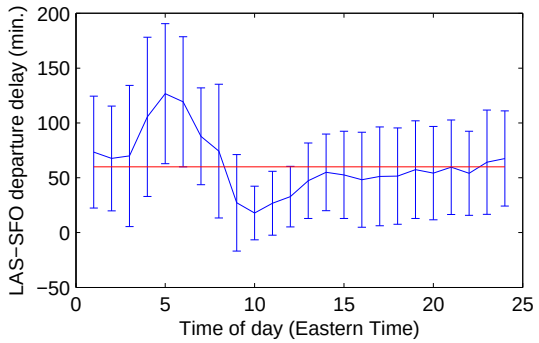

Fig. 12: LAS-SFO mean delay by time-of-day ($\pm\sigma$).

Figure 13 shows the histogram of the median test error for the 100 links studied. The median error values range from 15.6 min (EWR-ATL) to 36.4 min (LAX-HNL), and the average median test error is 20.9 min. As we can see in Figure 14 the standard deviation of these error values is low, the 90th percentile of the distribution is 1.17 min. It is remarkable that there is a gap between the highest median error value (LAX-HNL), and the second highest (SFO-JFK). Noting that none of these links had the highest test error in classification, we can ask the question: do links with high classification test error also have high regression test error? To answer this question, we plot the classification error versus the regression error (Figure 15). Although there is a strong positive correlation (0.78), some specific links perform significantly differently in the classification and regression problems. The highlighted data point in Figure 15 corresponds to the CLT-LGA departure delay prediction model. The classification test error in this link is 22.6%, which is high and in the 87th percentile of the classification error distribution, but the regression median test error is only 20.2 minutes, which is in the 40th percentile of the regression error distribution. This shows that a good performance in the regression problem does not necessarily mean good performance in the classification problem, and vice versa. The problems are different: in the classification problem we need information to allow us differentiate between high and low delay, but in the regression problem we need information to predict the *value* of the future delay. For a given link, it may be easier to predict if the future delay will be over 60 min, than to predict its exact value.

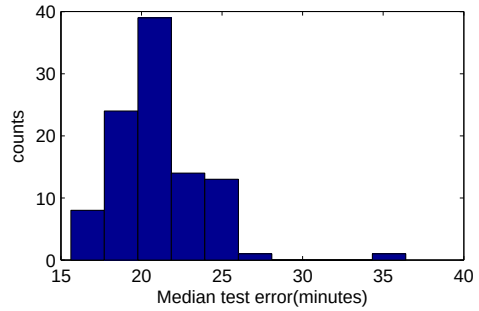

Fig. 13: Regression median test error histogram.

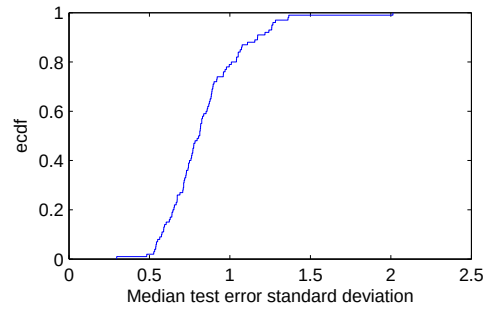

Fig. 14: Empirical cdf of the standard deviation of the regression median test error.

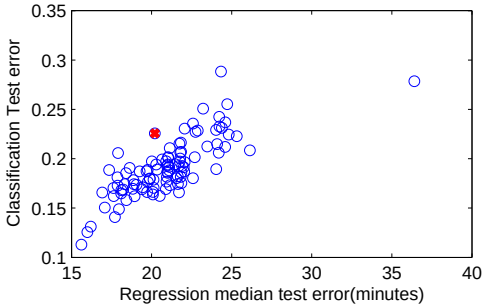

Fig. 15: Classification vs. regression test error.

C. Importance of explanatory variables

There are 26 explanatory variables in the prediction model. In this section, we evaluate the role of each of them by using the RF variables' importance to compare the relevance of the different variables. Table IV presents the average importance values for the 100 links. These results are for the 60-min threshold, 2-hour horizon classification problem.

TABLE IV
AVERAGE IMPORTANCE VALUES OF THE 100 OD PAIRS IN CLASSIFYING LINK DELAYS.

Variable	Var. Imp.	Variable	Var. Imp.
Time of day	78.5	NAS type of day	28.5
Day of week	6	NAS prev. type of day	18
Month of year	3	Top 3 airports average imp.	49.3
NAS delay state	19.1	Top 3 links average imp.	62.6

On average, we find that the time-of-day is the most important variable followed by the average importance of the three most important links. However, some links show very different behavior: For example, in predicting the ORD-PHL departure delay, the second most important variable after the current ORD-PHL departure delay is the NAS type of day, with an importance level of 72.2.

D. Effect of classification threshold

We also study the impact of changes in the classification threshold on the performance of the prediction models. We test three classification thresholds: 45, 60, and 90 min. The prediction time horizon continues to be 2 hours.

For the 100 most delayed links and the 45 minute threshold, we obtain a mean test error of 21.2%; for the 60 minute threshold, the misclassification test error is 19.1%; and for the 90 minute threshold,16.38%. The test error decreases as the classification threshold increases, since there are clearer indications of whether the future delay will exceed 90 min, than exceed 45 min.

Next, we look in more detail at the values of the test error for the 100 links studied. Figure 16 depicts the test error values for the three thresholds and the 100 links; the links are ordered according to their 60 min threshold test error. This plot shows that not all links have the same error reduction when increasing the classification threshold, and that this reduction is not correlated with the value of the test error. Figure 17 depicts the histogram of the test error increase when moving from a 90 min threshold to a 45 min threshold. For most links the error increases by 5 perc. points; however, the increase ranges from as low as 2 perc. points to 8 perc. points.

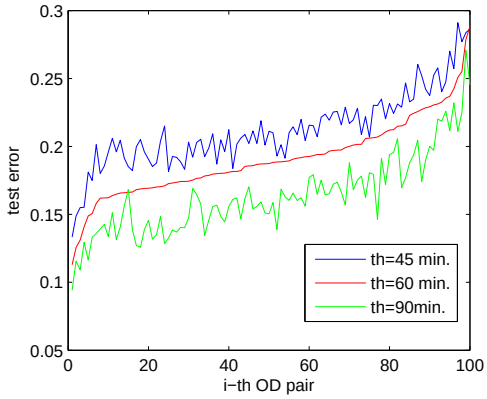

Fig. 16: Classification threshold analysis.

E. Effect of prediction horizon

One would expect the length of the prediction horizon to affect the prediction performance. We measure the impact of the prediction horizon length on the classification and regression problems. We analyze four different time horizons: 2, 4, 6 and 24 hours. The classification threshold is 60 min.

The average classification test errors for the 100 links and different time horizons are the following: 19.1% (2h), 21.4% (4h), 22.6% (6h), and 27.2% (24h). The average test error increase from 2 to 6 hours is only 3.5 percentage points. If we calculate the average test error for a model in which

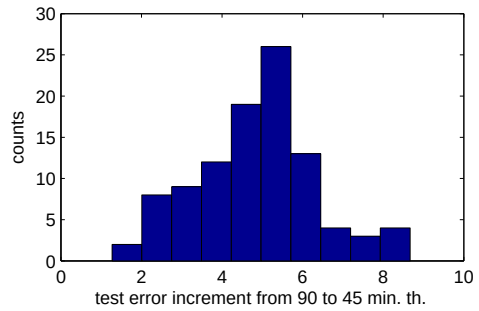

Fig. 17: Histogram of the test error increment when changing the classification threshold from 90 min to 45 min.

the only explanatory variable is the time-of-day, we have an average test error of 30%. The difference between this test error and the 24-hour horizon model test error is mostly due to the predictive value of the previous day's delay information. Figure 18 shows the test error values for the 100 links ordered in increasing order according to the 2h horizon test error. There is no correlation between the 2-hour horizon test error and the error increase as we increase the prediction horizon length.

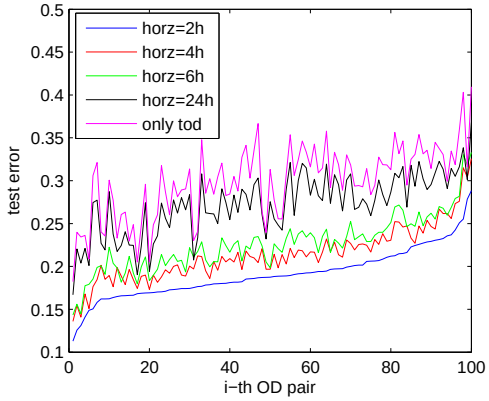

Fig. 18: Classification prediction horizon analysis.

Finally, we present the regression problem results. The average median test error for the 100 links and the different time horizons are the following: 20 min (2h), 23 min (4h), 24.3 min (6h), and 27.4 min (24h). In other words, the average median test error increase from 2 to 6 hours is only 4.3 minutes, and only 7.4 min as the prediction horizon increases from 2 to 24 hours.

VII. Conclusions

This paper presented a new network-based air traffic delay prediction model that incorporated both temporal and network delay states as explanatory variables. The results obtained for the 100 most-delayed OD pairs in the NAS showed an average test error of 19% when classifying delays as above or below 60 min, at a 2-hour prediction horizon. The analysis also found that the dependence of individual link delays on the network state varied from link to link. The results quantified the effects of the classification threshold and the prediction horizon on the predictive performance of the models. Both the classification and regression models were found to be quite robust to increases in the prediction horizon: The median regression test error (averaged across the 100 OD pairs) only increased from 20 min to 27.4 min when the prediction horizon increased from 2 hours to 24 hours.

The NAS delay state variables proposed in this paper enabled the development of the above network-based delay prediction models. These variables could potentially be used in the development of a network delay prediction and analysis tool. Other next steps in this research include the clustering of OD pairs by the predictive power of different explanatory variables, with the goal of identifying links in the NAS that exhibit similar behavior.

References

[1] S. AhmadBeygi, A. Cohn, Y. Guan and P. Belobaba, *Analysis of the potential for delay propagation in passenger airline networks*, Journal of Air Transport Management, Vol. 14, Issue 5, September 2008, pp. 221-236.

[2] M. Jetzki, *The propagation of air transport delays in Europe*, Thesis, Department of Airport and Air Transportation Research, Aachen University, 2009.

[3] Y. Tu, M. O. Ball and W. S. Jank, *Estimating flight departure delay distributions - A statistical approach with long-term trend and short-term pattern*, American Statistical Association Journal, 2008, vol. 103, pp 112-125.

[4] R. Yao, W. Jiandong and X. Tao, *A flight delay prediction model with consideration of cross-flight plan awaiting resources*, ICACC, 2010.

[5] B. Sridhar and N. Chen, *Short term national airspace system delay prediction*, Journal of Guidance, Control, and Dynamics, Vol. 32 No. 2, 2009.

[6] N. Xu, K. B. Laskey, G. Donohue and C. H. Chen, *Estimation of Delay Propagation in the National Aviation System Using Bayesian Networks*. Proceedings of the 6th USA/Europe Air Traffic Management Research and Development Seminar, 2005.

[7] Klein, A., Kavoussi, S., Hickman, D., Simenauer, D., Phaneuf, M., and MacPhail, T., *Predicting Weather Impact on Air Traffic*. ICNS Conference, Herndon, VA, May 2007.

[8] Klein, A., Craun, C., Lee, R.S. *Airport delay prediction using weather-impacted traffic index (WITI) model*. Digital Avionics Systems Conference (DASC), 2010.

[9] J. Rice, *Mathematical Statistics and data analysis*, 3rd ed., Duxbury Press. 2006.

[10] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning*, 2nd ed., Springer 2009.

[11] G. Upton and I. Cook, *A Dictionary of Statistics*, 2nd ed., Oxford University Press 2008.